DOCUMENT RESUME

ED 409 366 TM 026 904

AUTHOR Schnipke, Deborah L.; Reese, Lynda M.
TITLE A Comparison of Testlet-Based Test Designs for Computerized Adaptive Testing.
PUB DATE Mar 97
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability; *Adaptive Testing; Algorithms; Comparative Analysis; *Computer Assisted Testing; Difficulty Level; *Estimation (Mathematics); Psychometrics; *Scoring; Simulation; *Test Construction; Testing Problems
IDENTIFIERS *Testlets; *Two Stage Testing

ABSTRACT

Two-stage and multistage test designs provide a way of roughly adapting item difficulty to test-taker ability. All test takers take a parallel stage-one test, and, based on their scores, they are routed to tests of different difficulty levels in subsequent stages. These designs provide some of the benefits of standard computerized adaptive testing (CAT), such as increased precision of ability estimates over a paper-and-pencil test design. The item selection and scoring algorithms in two-stage and multistage designs may also be easier for test takers and test score users to understand--an important feature for gaining public acceptance of new test designs. This study incorporates testlets (bundles of items) into two-stage and multistage designs, and compares the precision of the ability estimates derived from these designs with those derived from a standard CAT design and from paper-and-pencil test designs. For the group that was used to establish the cutoffs for the two-stage and multistage testlet designs, 50,000 simulated test takers were created randomly. The group of simulated test takers used to simulate all test designs totaled 25,000. Results indicate that all testlet-based designs resulted in improved precision over the same-length paper-and-pencil test, and almost as much precision as the paper-and-pencil test of double length. Given the many other (nonpsychometric) advantages of these designs, they may be viable options for computer-administered tests. (Contains six figures and nine references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Deborah Schnipke
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

A Comparison of Testlet-Based Test Designs for Computerized Adaptive Testing[1]

Deborah L. Schnipke and Lynda M. Reese[2]
Law School Admission Council

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

ED 409 366

Abstract: Two-stage and multistage test designs provide a way of roughly adapting item difficulty to test-taker ability. All test takers take a parallel stage-one test, and, based on their score, they are routed to tests of different difficulty levels in subsequent stages. These designs provide some of the benefits of standard computerized adaptive testing (CAT), such as increased precision of ability estimates over a paper-and-pencil test design. Additionally, the item selection and scoring algorithms in two-stage and multistage designs may be easier for test takers and test-score users to understand – an important feature for gaining public acceptance of new test designs. This study incorporates testlets (bundles of items) into two-stage and multistage designs, and compares the precision of the ability estimates derived from these designs with those derived from a standard CAT design and from paper-and-pencil test designs. Results indicate that all testlet-based designs resulted in improved precision over the same-length paper-and-pencil test, and almost as much precision as the paper-and-pencil test of double length. Given the many other (nonpsychometric) advantages of these designs, they may be viable options for computer-administered tests, and future research will continue to investigate these designs.

Because of the many benefits of computer-administered testing (e.g., the potential for new item types, more frequent testing, immediate scoring), the Law School Admission Council (LSAC) is considering computerizing the Law School Admission Test (LSAT). Several concerns have been raised about the standard maximum-likelihood computerized adaptive test (CAT) design for the LSAT, however. For example, it would be difficult to explain information-based item selection and maximum-likelihood or Bayes-modal scoring to test takers and test-score users. While the LSAC is interested in a computerized LSAT that adapts item difficulty to test-taker ability, we are interested in investigating less complicated (and easier to explain) ways of doing so.

One possible way to simplify the adaptive nature of a CAT is to use a two-stage test design (Lord, 1971, 1980). The concept of two-stage testing emerged as a rudimentary means of tailoring the difficulty level of a test to the ability level of a test taker before advances in computer technology made CAT feasible. In a two-stage design, test takers first take a "routing test" of

TM026904

[1]The authors wish to acknowledge the programming support of Jennifer Lawlor.

[2]This research was collaborative in every respect, and the authorship is shared equally. This paper was presented at the annual meeting of the American Educational Research Association, Chicago, March, 1997.

BEST COPY AVAILABLE

average difficulty (stage one). Based on their number-right scores on the routing test, test takers are routed to a "measurement test" that is roughly adapted to their ability level (stage two). Test takers with low scores on the routing test are administered an easier test in the second stage, test takers with high scores are administered a more difficult second-stage test, and test takers with middle scores receive another average-difficulty test. In this way, item difficulty is roughly adapted to test-taker ability in the second stage. Lord (1971) showed that a two-stage design provides more precise measurement than an equal length nonbranching test. Lam and Foong (1991) also noted greater precision for two-stage testing as compared to a linear paper-and-pencil test.

The two-stage design can be expanded to a multistage design where test takers are routed to more narrowly focused tests at higher stages. For example, test takers can be routed from the first stage to a low, medium, or high stage two level, and next to a very low, low, medium, high, or very high stage three level. Such designs may be appropriate for a computer-administered version of the LSAT given the perceived need to explain item selection to test takers and users of test scores.

Additional future concerns will include making provisions for items that refer to a common stimulus (set-bound items, such as reading comprehension) and whether to allow item review (e.g., see Stocking, 1996). LSAC field tests have shown that test takers strongly desire the capability of reviewing/revising previous responses (S. Jenkins, personal communication, July 24, 1996), and this capability will be incorporated if possible. The use of testlets (bundles of items which are administered as a unit) may provide a solution to these concerns (Wainer & Kiely, 1987). A common stimulus (e.g., a reading passage) and its associated items can be designated as a testlet; thus, the items will automatically be administered together. By not adapting within a testlet, item review within a testlet can be allowed without leading to undesirable test-taking strategies affecting the precision of the test. Results from our field tests indicate that test takers are comfortable with item review within a testlet.

The two-stage or multistage test can be built from testlets, and such a design may provide a solution to the concerns raised about the standard CAT design for the LSAT. The present study uses simulated data, based on simulated test taker and item parameters, to determine the precision of ability estimates of various test designs. The test designs, described in more detail below, are a two-stage testlet design, a two-stage testlet design that reroutes test takers within the second stage as needed, a multistage testlet design (which had four stages in the present study), a standard

maximum-information item-level design (e.g., Wainer, et al., 1990, which is the psychometric ideal in terms of precision and efficiency), a maximum-information testlet-based design (which adapts at the testlet level rather than the item level), and a paper-and-pencil (i.e., nonadaptive) design of two lengths (the same length as the other designs and twice as long). The paper-and-pencil design of the same length as the other designs serves as the minimally acceptable criterion for the new designs.

Method

Simulated test takers

Two groups of simulated test takers were created. One group was used to establish the cutoffs for the two-stage and multistage testlet designs (described below), and the other group was used for the simulations of all test designs. For the group that was used to establish the cutoffs for the two-stage and multistage testlet designs, 50,000 simulated test takers were created by randomly sampling ability (θ) parameters from a standard normal distribution. A normal distribution was used so that we could track the number of test takers in a typical population who would be routed to the various levels.

The group of simulated test takers used to simulate all test designs was defined as 1,000 θ's from -3 to 3 in increments of 0.25, for a total of 25,000 simulated test takers. This flat distribution of ability values was used so that the precision of ability estimates across the entire ability range could be determined accurately.

Item Parameters

Testlets were created specifically for the two-stage and multistage designs. The testlets for the multistage design were also used for the maximum-information testlet-based design. The items that comprised the testlets for the multistage design were used for the standard maximum-information item-level design.

For each stage/level, item parameters were generated for one testlet at a time, beginning with the *b*-parameter. *B*-parameter values were generated for a five-item testlet by selecting randomly from a normal distribution with the specified mean and a standard deviation of 0.8. The mean *b* for stage-one testlets was set at -0.5. Stage-two testlets were centered at b=-1.0 (low), b=0.0 (medium), or b=1.0 (high). (Stage-one and stage-two testlets were identical for the two-stage and multistage

designs.) Stage-three testlets for the multistage design were centered at b = -1.25, -.75, .75, or 1.25. Stage-four testlets for the multistage design were centered at b = -1.5, -1.0, 0.0, 1.0, or 1.5. Note that stage one had one level, stage two had three levels, stage three had four levels, and stage four had five levels. A testlet for any stage/level was retained only if the difference between the lowest and highest b-parameter value for that testlet was between 1.5 and 2.0 and if the mean of the b values for that testlet was within .3 of the specified mean. Any testlet that did not meet these requirements was rejected. This insured that the testlets within a given stage/level would have b values that were comparable across testlets, creating testlets that were essentially parallel to one another in terms of difficulty.

Once the b-parameter values were generated satisfactorily, a- and c-parameter values were generated. The a's for stage one were drawn from a normal distribution with a mean of 0.8 and a standard deviation of 0.22. The a's for stages two, three, and four were drawn from a normal distribution with a mean of 0.9 and a standard deviation of 0.22. The a's for stage one were lower on average because we wanted to save the "better" (more discriminating) items for later stages when item difficulty was closer to test-taker ability. The c's for all stages/levels were drawn from a uniform distribution ranging from 0.15 to 0.25, which is roughly comparable to five-option, multiple-choice items.

Test Designs

Two-Stage Testlet Design

In the two-stage testlet design, the number-right score on stage one was used to route test takers to stage two, where item difficulty more closely matched test-taker ability (e.g., more difficult testlets for higher ability test takers). In stage one, two testlets were randomly selected for each simulated test taker in the group from the flat-ability distribution. The simulated test taker's number-right score was calculated and was used to route the simulated test taker to a low, medium, or high stage two level. As shown in Figure 1, stage-one number-right scores of 0 to 6 were routed to low, 7 to 8 were routed to medium, and 9 to 10 were routed to high stage-two testlets. (How we determined which scores to route to the various levels is discussed below.) In stage two, three testlets were randomly selected at the appropriate level (low, medium, or high, based on the stage-one number-right score) and administered to each simulated test taker.

After all items were administered, the final θ estimate was calculated using Bayes modal scoring (Hambleton, Swaminathan & Rogers, 1991), based on all 25 responses, with a standard normal prior distribution for θ. (Bayes modal scoring requires an initial θ estimate. The initial estimate was obtained with Owen's Bayes sequential scoring (Owen, 1969) which updated the θ estimate after each item was administered.)

Routing in Stage Two Based on Stage One Number-Right Score

The purpose of stage two is to tailor item difficulty more closely to test-taker ability. Matching item difficulty to test-taker ability will maximally decrease measurement error for a fixed number of items. Thus, the level (low, medium, or high) of stage two which is expected to decrease measurement error the most for a given test taker is the one that should be administered to that test taker. For each number-right score, the error that would result if each level of stage two was administered separately was determined.

Specifically, the mean-squared error (MSE) of ability (θ) was used to determine which stage-one number-right scores would be routed to each level (low, medium, or high) of stage two. Test takers with a given stage one number-right score should be routed to the stage two level that leads to the lowest MSE for that number-right score. To determine the cutoff scores for routing simulated test takers to stage-two levels of low, medium and high, all simulated test takers in the sample of 50,000 simulated test takers with a standard normal ability distribution were first administered two stage-one, five-item testlets, and their number-right score was calculated. Regardless of stage-one number-right score, each simulated test taker was administered three randomly selected low stage-two testlets, and θ estimates were obtained using all stage-one and stage-two items administered. All simulated test takers were next administered three randomly selected medium stage-two testlets, and new θ estimates were obtained using the responses to items on stage one and the medium stage-two testlets that were administered to the test taker. Finally, all simulated test takers were administered three randomly selected high stage-two testlets, and a third θ estimate was obtained for each test taker using stage one and the high stage-two testlets.

MSE_s was calculated separately for the three θ estimates (one from each level of stage two) at each stage-one number-right score, s. MSE_s is given by

$$MSE_s = \frac{1}{N}\sum_{i=1}^{N}(\theta_i - \hat{\theta}_i)^2$$

where θ_i represents the true value of the ability parameter for test taker *i*,

$\hat{\theta}_i$ represents the estimated ability value for test taker *i*, and

N is the number of simulated test takers who obtained a number-right score of *s*.

Figure 2 shows the MSE_s values for the two-stage design. Test takers who obtained a low stage-one number-right score (less than 6 correct) presumably have a low true θ, and a low stage-two testlet leads to the lowest measurement error (MSE). Similarly, test takers who obtained a high stage-one number-right score (9 or more correct) presumably have a relatively high true θ, and a high stage-two testlet leads to the lowest measurement error (MSE). The locations at which the low and medium and the medium and high lines crossed determined the stage-one number-right cutoffs between the low, medium, and high levels. In this case, stage-one number-right scores of 0-6 were routed to low, 7-8 were routed to medium, and 9-10 were routed to high stage-two testlets.

Two-Stage Testlet Design with Changing Levels

As a variation on the two-stage design, we repeated the two-stage design with the exception that simulated test takers were rerouted to a different level within stage two if they were determined to be misclassified. Number correct cutoffs scores for routing simulated test takers to alternate stage-two levels were determined in the same manner described above for determining routing after the stage-one test. After being routed to a stage-two level and being administered an initial stage-two testlet, all simulated test takers were administered one subsequent testlet from each of the stage-two levels, and a θ estimate was derived for each stage-two level for each simulated test taker. MSE_s was calculated as described above separately for each of the three θ estimates for each simulated test taker within each stage-two level. These values were plotted separately for the low, medium and high stage-two levels, and the locations where the lines crossed determined the number-right cutoffs for reclassification after the initial stage-two testlet. This same analysis was

carried out after the second stage-two testlet was administered. The number-right scores that were routed (and rerouted) to (and within) stage two are shown in Figure 3.

As in the original two-stage design, the standard normal group of 50,000 θ's was used to determine the cutoffs. The group of 25,000 θ's with the flat distribution was used for the actual simulations. The final ability estimate was the Bayes modal estimate, based on all 25 items administered.

Multistage Testlet Design

As another variation on the two-stage design, we created a multistage design. In this design, all test takers received two stage-one testlets centered at b=-0.5, one stage-two testlet centered at b=-1.0, 0.0, or 1.0, one stage-three testlet centered at b=-1.25, -0.75, 0.75, or 1.25, and one stage-four testlet centered at b=-1.5, -1.0, 0, 1.0, or 1.5. As in the two-stage design, test takers were routed to the various levels based on their number-right score on the previous stage. As in the two-stage design, the cutoffs were established by calculating the error (MSE_s) that would result in the ability estimate if simulated test takers in a given level with a given number-right score were administered each of the levels of the next stage.

The standard normal group of 50,000 θ's was used to determine the cutoffs. The number-right scores that were routed to each stage/level are shown in Figure 4. The group of 25,000 θ's with the flat distribution was used for the actual simulations. The final ability estimate was the Bayes modal estimate, based on all 25 items administered.

Standard Maximum-Information Item-Level Design

Item selection for the standard maximum-information item-level design was based on item information, as specified by item response theory. Item information, $I_i(\theta)$ was calculated at 37 θ values (from -2.25 to 2.25 in increments of 0.125) for each item using the formula

$$I_i(\theta_j)=\frac{2.89a_i^2(1-c_i)}{[c_i+e^{1.7a_i(\theta_j-b_i)}][1+e^{-1.7a_i(\theta_j-b_i)}]^2}$$

where i indicates the item,

j indicates the θ value,

a_i is the IRT discrimination parameter for item i,

b_i is the IRT difficulty parameter for item i, and

c_i is the IRT lower asymptote parameter for item i (Hambleton, Swaminathan, & Rogers, 1991).

The information values were used during the simulations to select the items with the highest information at a given θ-level. The flat distribution of 25,000 simulated test takers was used for the simulations. A fixed-length, 25-item CAT was simulated.

To prevent items from becoming "overexposed" (administered to too large a proportion of simulated test takers), a 10-9-8- . . . exposure-control method (Kingsbury & Zara, 1989) was incorporated into the simulations. The first item to be administered to a simulated test taker was randomly selected from the 10 items with the highest information values at θ=0 (the starting value for all simulated test takers). The second item was randomly selected from the 9 best items at the new estimate of θ. The third item was randomly selected from the 8 best items, and so on until, beginning with the tenth item, the item with the highest information was selected (unless, of course, the item had already been administered to that simulated test taker, in which case the next best item was selected).

After each item was selected, the simulated test taker's response (right/wrong) was determined, and the simulated test taker's estimated θ was updated using Owen's Bayes sequential scoring (Owen, 1969). After all items were administered, a Bayes modal score (e.g., Hambleton, Swaminathan, & Rogers, 1991) was calculated and was used as the final theta estimate.

Maximum-Information Testlet-Based Design

Item selection for the maximum-information testlet-based design was based on testlet information. Item information was calculated at 37 θ values (from -2.25 to 2.25 in increments of 0.125) for each item and was summed across items in the testlet to indicate testlet information.

The testlet information values were used during the simulations to select the testlets with the highest information at a given θ-level. A fixed-length 25-item test was simulated for each simulated test taker in the group from the flat-ability distribution.

To prevent testlets from becoming "overexposed" (administered to too large a proportion of simulated test takers), a 10-9-8- . . . exposure-control method (Kingsbury & Zara, 1989) was

incorporated into the simulations. The first testlet to be administered to a simulated test taker was randomly selected from the 10 testlets with the highest information values at $\theta=0$ (the starting value for all simulated test takers). The second testlet was randomly selected from the 9 best testlets at the new estimate of θ. The third testlet was randomly selected from the 8 best testlets, and so on until all 5 testlets were selected.

After each testlet was selected, the simulated test taker's response (right/wrong) was determined for each item in that testlet, then the simulated test taker's estimated θ was updated using Owen's Bayes sequential scoring (Owen, 1969). After all items were administered, a Bayes modal score (e.g., Hambleton, Swaminathan, & Rogers, 1991) was calculated and was used as the final θ estimate.

Paper and Pencil Design

The items in the paper-and-pencil designs were taken from two intact LSAT test sections which are designed provide the best measurement in the middle of the ability distribution where the bulk of the test takers is in a typical test-taker population. The sections had 25 and 26 items. We simulated responses (using the flat distribution of 25,000 θ values) for both sections and used the 25-item section and both sections combined (51 items) in subsequent analyses.

Analyses

To indicate the amount of error in the ability estimates, the root mean squared error (RMSE) was plotted for each test design at each θ level. To indicate whether ability is overestimated or underestimated, the bias statistic was also plotted. Positive bias values indicate that ability was underestimated, and negative values indicate ability was overestimated.

The root mean squared error (RMSE) is given by

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\theta-\hat{\theta})^2}.$$

The bias statistic is given by

$$Bias = \frac{1}{N} \sum_{i=1}^{N} (\theta - \hat{\theta}).$$

Results

As shown in Figures 5 and 6, the standard maximum-information item-level CAT (which adapts at the item level rather than the testlet level) led to less error in ability estimates (smaller RMSE) and less bias, particularly in the tails of the ability distribution, than any of the other designs. This is not surprising since the standard CAT design adapts the difficulty of the item to the test taker's estimated ability after every item, rather than after blocks of 5 or 10 items as in the other adaptive designs. Although this design is not practical for a computerized version of the LSAT, it does provide a lower limit on how little error there could be in the ability estimates produced by the other designs.

The 25-item paper-and-pencil design led to the most error (RMSE) and bias in ability estimates. The two-stage, multistage, and maximum-information testlet-based designs (which were all 25 items in length) led to ability estimates that were very similar in terms of RMSE and bias to the 51-item paper-and-pencil design for θ's less than 1.5. For θ's greater than 1.5, the 51-item paper-and-pencil design led to θ's with less error and less bias than the two-stage and multistage designs. The two-stage and multistage designs did lead to θ's with less error and bias than the equal length (25-item) paper-and-pencil design for all θ's, including those greater than 1.5. The maximum-information testlet-based design led to θ's that had slightly less error and bias than the 51-item paper-and-pencil design, especially in the tails of the ability distribution.

Discussion

The standard maximum information CAT design is impractical for many large-scale testing programs because of nonpsychometric considerations. For instance, items that refer to a common stimulus must be administered together. Additionally it would be advantageous from the test takers' perspective to allow item review. Although these considerations are possible to accommodate in a maximum-information CAT, testlets may provide a solution to these and other considerations.

In terms of RMSE and bias, all testlet-based designs resulted in improved precision over the same-length paper-and-pencil test, and almost as much precision as the paper-and-pencil test of double length. The two-stage and multistage designs were very similar to each other across the entire ability scale. In terms of psychometric characteristics, the two-stage and multistage designs performed at an acceptable level. Given the many other (nonpsychometric) advantages of these designs, they may be viable options for a computerized LSAT, and future research will continue to investigate these designs.

References

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Lam, T. L. & Foong, Y. Y. (1991, April). *Development and evaluation of hierarchical testlets in two-stage tests using integer linear programming*. Paper presented at the meeting of the American Educational Research Association, Chicago.

Kingsbury, G. G. & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*, 359-375.

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika, 36*, 227-242.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Owen, R. J. (1969) *Bayesian approach to tailored testing*. (Research Bulletin 69-92). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1996). *Revising answers to items in computerized adaptive tests: A comparison of three models* (Research Report 96-12). Princeton, NJ: Educational Testing Service.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

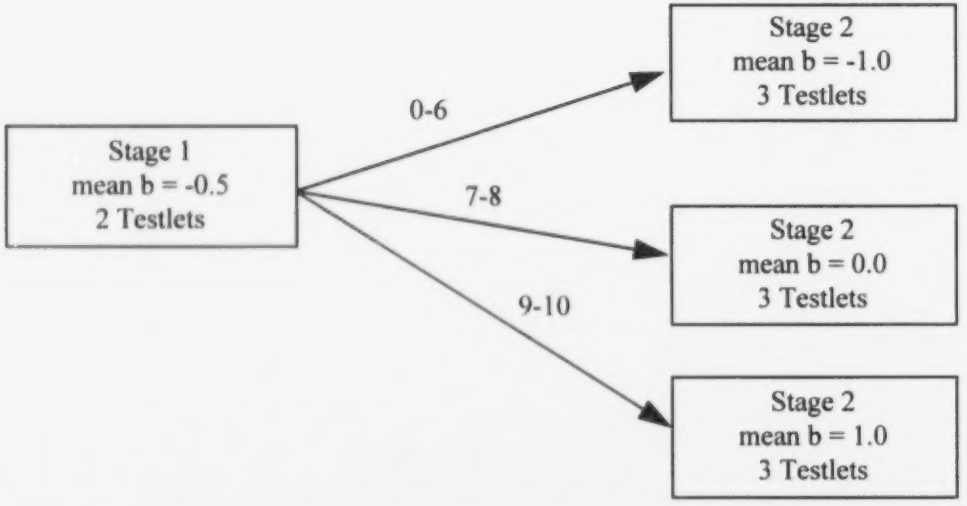

Figure 1. Stage-one number-right scores that are routed to each level of stage two.

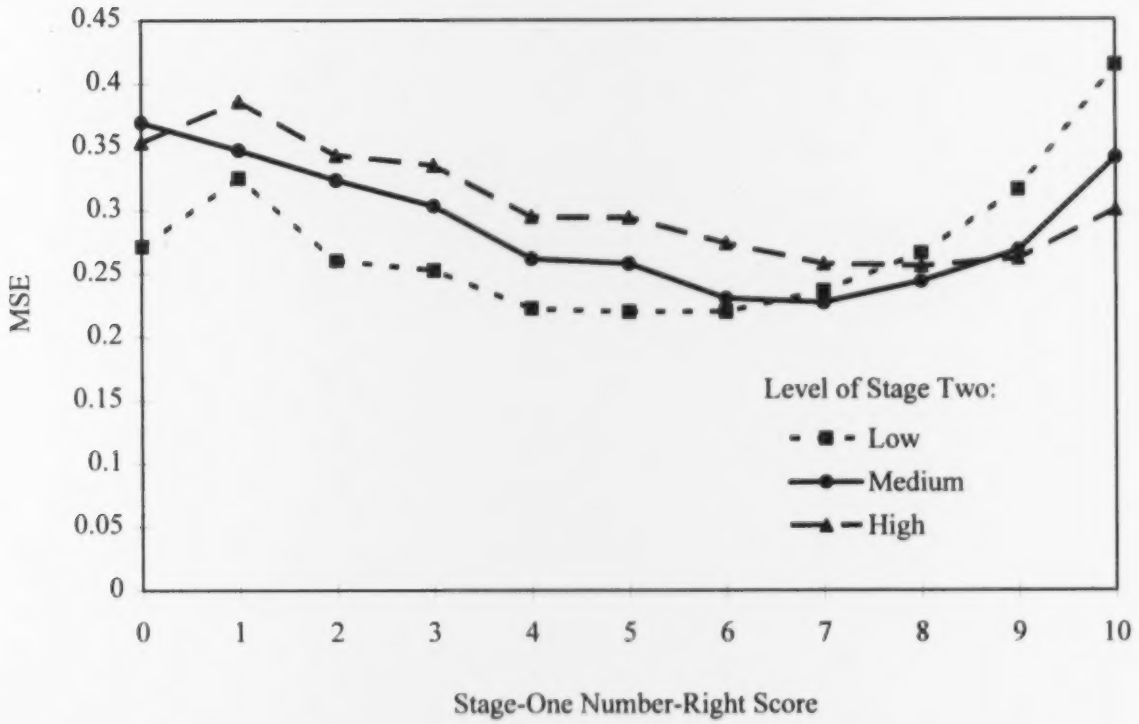

Figure 2. MSE values for the two-stage design. MSE indicates the amount of error there would be in ability estimates if each level of stage two were administered to examinees with a given stage-one number-right score.

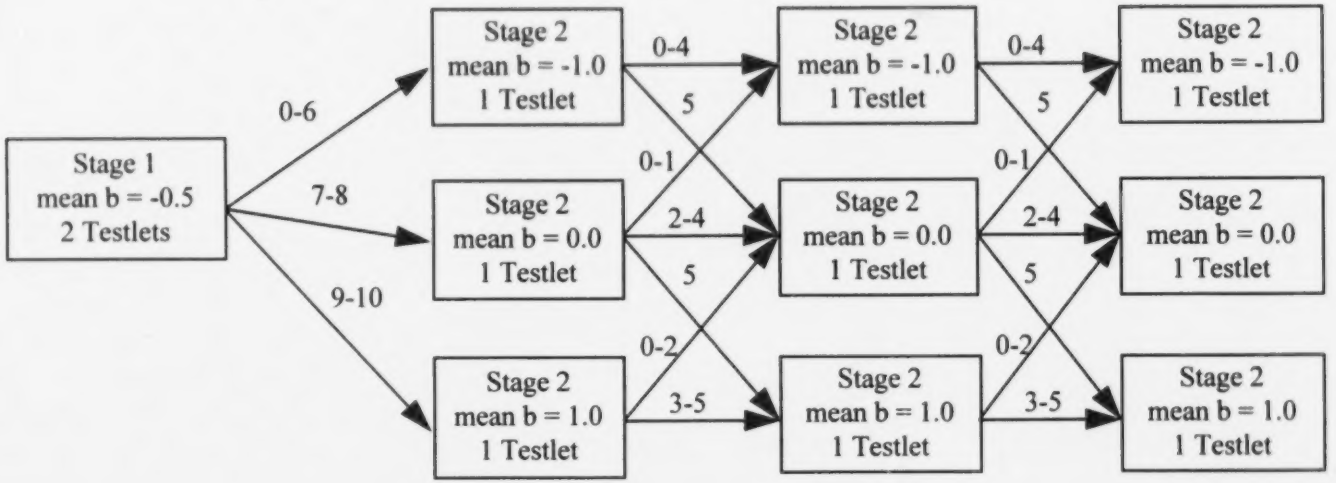

Figure 3. Number-right scores that are routed to each level of stage two. Examinees are rerouted in stage 2 after each testlet.

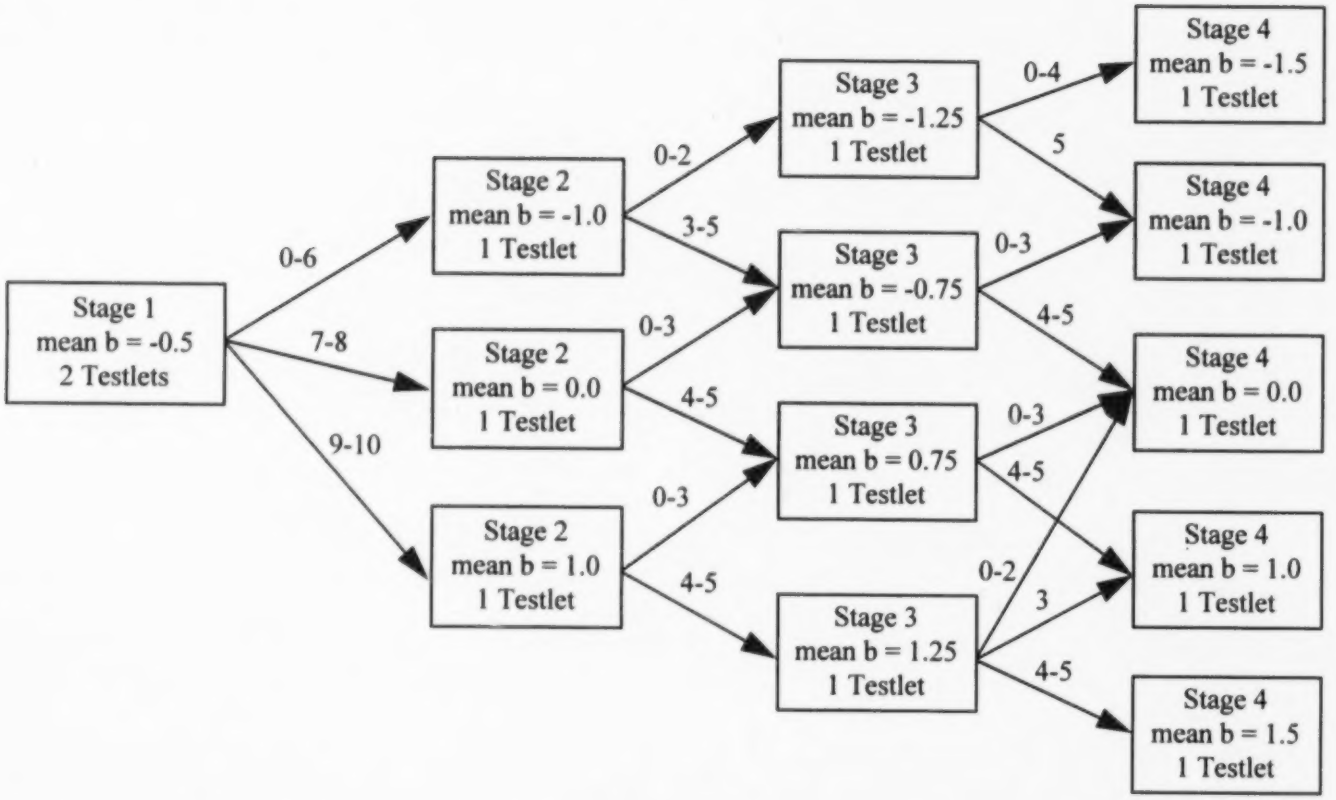

Figure 4. Multistage Design. Number-right scores that are routed to each level of the next stage are indicated above each arrow.

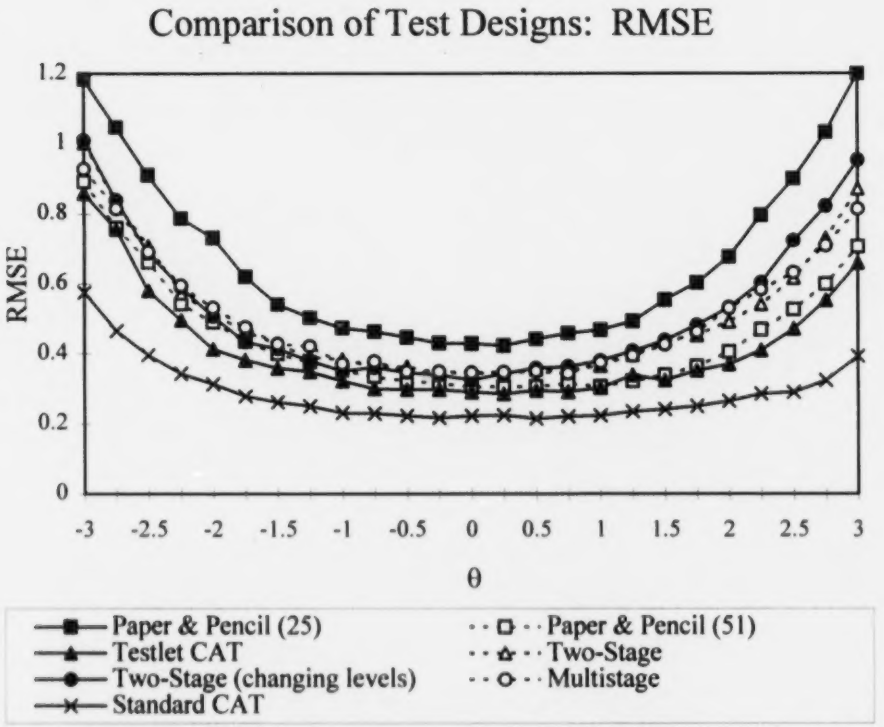

Figure 5. RMSE values for each test design, calculated at each θ level.

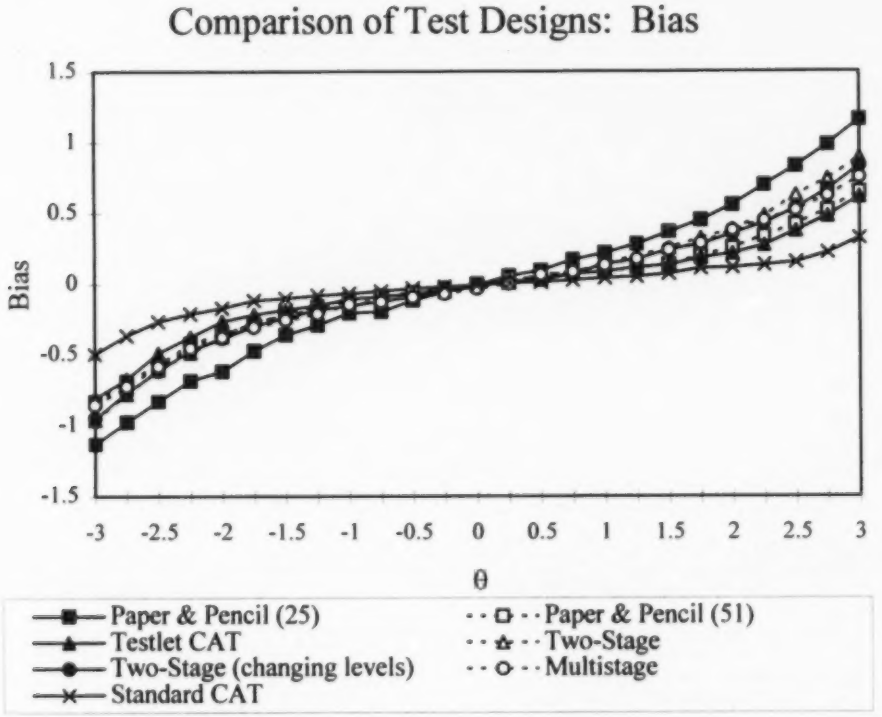

Figure 6. Bias values for each test design, calculated at each θ level.